Annotating Genericity: How Do Humans Decide? (A Case Study in Ontology Extraction)

Aurelie Herbelot and Ann Copestake

University of Cambridge

`ah433@cam.ac.uk`

1 Introduction

In computational linguistics, the task of ontology extraction deals with acquiring facts from natural language for inclusion in knowledge bases (so-called 'ontologies'). Recognising genericity plays an important role in ensuring that the extracted information is correct. For instance, let's consider the following two sentences, taken from a single Wikipedia article:

> The whale feeds mainly on benthic crustaceans [...]. Mostly, the animal feeds in the northern waters during the summer.

An ontology extraction system would have to understand that the first sentence relates to the kind 'whale' and not to a single individual. In the second sentence, the anaphoric relation requires the resolution of 'animal' to the kind 'whale' to avoid incorrect conclusions such as 'all/most animals feed in the northern waters.'

Our project involves training a machine learner on a large set of examples annotated by humans. Unfortunately, for genericity, no adequate data set is readily available and there are no standard guidelines spelling out the decision process that a human goes through when understanding a particular NP as specific or generic. The annotation schemes we know of (ACE, 2005 or Behrens, 2005 for instance) follow simple guidelines which, unfortunately, are not suitable for our purposes (for some criticisms of the ACE guidelines, see Suh, 2006, Section 3.2.4). This paper therefore discusses a more thorough scheme adapted to the task of ontology extraction.

In the next sections, we introduce our scheme and compare six separate sets of manual annotations produced for a given Wikipedia article. The evidence obtained when considering disagreements allows us to point out the main areas of difficulty experienced by humans when dealing with the genericity phenomenon.

2 Manually Annotating Genericity

Our annotation scheme is geared towards the task of ontology extraction. We started from the assumption that the facts expressed in an ontology should be true when taken out of textual context. That is, given a network of concepts linked by relations, the scope of concept nodes should clearly be indicated in the network. We believe that this is not only a matter of labelling specific (non-kind) against generic (kind) entities and we therefore propose a larger set of annotations:

- Noun phrases which are in appearance generic do not necessarily refer to all instantiations – or to a stereotypical instantiation – of the concept they lexically express (see the 'animal' relation in the introduction). We therefore use a GROUP label to denote generics which only refer to a defined subgroup of the concept expressed by the lexical string.[1]

- Carlson et al (1995) argue that the notion of genericity potentially only applies to 'well-established' entities. We follow this argument because of its ontological relevance (not all noun phrases are suited for inclusion in an ontology.) We use a NON-WD label to denote entities which are not well-defined and would not appear in an ontology (because there is no clear definition of 'well-established', we rely on the annotators' intuition of what makes a clear concept, as shown below.)

We give next a summarised version of our annotation scheme:

(1) Can the entity be singled out from similar entities (that is, entities expressed by the same lexical string) in the real world? If yes, annotate as SPEC.

(2) Is it possible to imagine instantiations or subgroups of the expression? If not, annotate as SPEC.

(3) Is the expression a 'well-defined entity'? (Could you imagine an encyclopedia article or web page on the topic?) If not, annotate as NON-WD.

(4) Can the information contained in the text apply to the entity in general (that is, to all or any of its instances), or is it only relevant to a subgroup? If general, annotate as GEN, otherwise, annotate as GROUP.

3 Annotation Analysis

The scheme was evaluated by presenting six annotators with four sections of the Wikipedia article 'The History of Cryptography', totalling 100 noun phrases, the boundaries of which were marked by one of the authors prior to annotation. The evaluation gave us 15 kappa figures ranging from 0.31 to 0.62, with agreement ranging from 56% to 78%. We carried out an analysis of the areas of stronger

---

[1] This does not only concern anaphora resolution. Consider, for instance, a book about 19th-century Europe and the sentence 'Women were not allowed to vote.' The scope of the NP 'women' is 19th-century Europe and it would be incorrect to generically relate the information to all women.

disagreement by picking out all NPs without majority. That is, we selected all phrases for which fewer than four annotators agreed on a label. We then attempted to understand the reason for each annotation and derived a rough map of issues.

We found that for 72 phrases out of 100, at least four annotators were in agreement. The remaining 28 phrases formed the core of our analysis. The following conclusions could be derived:

- A large portion of the disagreements could have been covered by an 'ambiguity' label. In those cases, there is a genuine ambiguity between specificity and genericity, which is not resolved in the text. For instance, in the sentence *'Later still, Hebrew scholars made use of simple mono-alphabetic substitution ciphers'*, it is not clear whether all Hebrew scholars or some specific people are involved.

- Our simple definition of genericity in step 4 (*all* or *any* of the instances of the concept expressed by the lexical item) produces confusions between GROUP and GEN labels. In the sentence *'The Greeks of classical times are said to have known of ciphers'*, 'ciphers' was labelled by some annotators as GROUP, presumably because the Greeks would not have known of all ciphers in existence.

- The semantic interpretation of the copula differs depending on the annotator. Some assume that the second argument of the copula has automatically the same scope as the first argument, while others prefer to see it as a class (and therefore a generic entity). An example of such disagreement occurs for the NP 'examples of cryptography' in the sentence *'Herodotus tells us of secret messages [...] these are not proper examples of cryptography.'* Out of the two annotators having marked 'secret messages' as specific, one marked the last NP as generic, the other as specific.

- In some cases, the disagreement is simply due to slight differences in the resolution of the referent in context.

Having noted these points, we have now modified the annotation scheme to include an ambiguity label, hard instructions regarding the copula, a refined way to get to the GEN label and instructions regarding referent resolution. The revised guidelines will be used in a new annotation experiment in the hope that inter-annotator agreement increases to the point of usability.

References

1. Behrens, L. *Genericity from a Cross-Linguistic Perspective.* In Linguistics, 43/2, 2005.
2. Carlson, G. and Pelletier, F. *The Generic Book.* University of Chicago Press, 1995.
3. Suh, S. *Extracting Generic Statements for the Semantic Web.* MSc thesis In Artificial Intelligence, School of Informatics, University of Edinburgh, 2006.
4. ACE (Automatic Content Extraction) English Annotation Guidelines for Entities, Version 5.6.1 2005.05.23, 2005.